KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Improved lexical similarities for hybrid clustering through the use of noun phrases extraction

Bart Thijs, Wolfgang Glänzel and Martin Meyer

MSI_1703

Improved lexical similarities for hybrid clustering through the use of noun phrases extraction

Bart Thijs[1], Wolfgang Glänzel[2], and Martin Meyer[3]

[1] *bart.thijs@kuleuven.be*
KU Leuven, FEB, ECOOM, Leuven (Belgium)

[2] *wolfgang.glanzel@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

[3] *m.s.meyer@kent.ac.uk*
Kent Business School, University of Kent, Canterbury(UK)
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
SC-Research, University of Vaasa, Lapua, (Finland)

Abstract

Clustering of hybrid document networks combining citation based links with lexical similarities suffered for a long time from the different properties of these underlying networks. In this paper we evaluate different processing options of noun phrases extracted from abstracts using natural language processing to improve the measurement of the lexical component. Term shingles of different length are created from each of the extracted noun phrases. We discuss twenty different extraction-shingling scenarios and compare their results. Some scenarios show no improvement compared with the previously used single term lexical approach used so far. But when all single term shingles are removed from the dataset the lexical network has properties which are comparable with those from a bibliographic coupling based network. Next, hybrid networks are built based on weighted combination of the two types of similarities with seven different weights. We demonstrate that removing all single term shingles provides the best results at the level of computational feasibility, comparability with bibliographic coupling and also in a community detection application.

Introduction

For a long time scientometrians have been using the combination of textual analyses with citation based links for many different applications. In 1991, Braam et al. (1991a & b) suggested the use of co-citation in combination with word-profiles which are indexing terms and classification codes for a mapping of science. In the same year, Callon et al. (1991) demonstrated how co-word analysis can be used for studying academic and technological research. Later, Noyons and Van Raan (1994) constructed geometrically organized maps based on co-occurrence of keywords in patents and publications to illustrate possible links between science and technology. In the same year Zitt and Bassecoulard (1994) used lexical and co-citation analysis for trend detection and analysis. In 2005, Glenisson et al. (2005a and 2005b) started using full text instead of the rather limited set of keywords originating from subject heading, titles or abstracts and they could compare the performance of the full text approach with the combined approach with title/abstract and reference-based text analysis. The application of bibliometric indicators allowed to fine-tune the clusters found after full text mining.

Processing these full texts introduced new problems that were less likely to occur with keyword based approaches. Stemming was required to reduce the English words to their stem (the Porter stemmer was used for this purpose, see Porter, 1980). The dimensionality of the representation of documents in a vector space grew and Singular Value Decomposition had to be introduced to reduce this dimensionality. Glenisson encountered the disadvantage of the single term

approach and used the Dunning likelihood ratio test (Dunning 1993; Manning & Schütze, 2000) to identify common bigrams. For this test the occurrence of each pair has to be calculated together with the frequency of each term appearing separately. The bigrams with the highest score are retained. The risk of this procedure is that pairs that are less frequent or that appear in a few variations are not selected. Also the selection of a bigram in a paper might change when additional documents are added to the dataset. It is clear that the introduction of full text analysis increased processing complexity.

Janssens (2005) introduced a true integrated approach where he combines the distance based on bibliometric features with a text-based distance using the following scheme:

$$D^{INTEGR} = \lambda D^{TEXT} + (1-\lambda) D^{BIBL}$$

While introducing this simple weighted linear combination, he immediately listed some issues that need to be solved. One of these is the different distributional characteristics of the combined similarities and another is the choice of the weighting parameter which is set manually and is quite arbitrary. In this 2005 paper the authors use a Silhouette Value per Cluster as introduced by Jain and Dubes (1988) as an estimator for λ. Later Janssens et al. (2008) warned against the combination based on simple vector concatenation and linear combinations of similarity measures because of the completely different structures of the underlying vector spaces and they proposed a combination based on Fisher's inverse Chi-Square. They also showed that this method outperforms hitherto applied methods. Each similarity is converted into their corresponding *p*-value based on a cumulative distribution function of the similarities in a completely randomized dataset. This method solves the issue of different distributions drastically but it introduces an even more complex calculation scheme. Glanzel & Thijs (2012) take a more pragmatic approach and exploit the fact that both similarities can be expressed as cosines in a vector space model and introduce a hybrid similarity as the cosine of the weighted linear combination of the underlying angles of each of the cosine similarities.
None of solutions proposed in the literature were so far able to eliminate or at least to considerably reduce the effect of different distributions in each of the two components without excessive computational requirements.

In this paper we introduce the use of noun phrase extracted by the application of Natural Language Processing (NLP) and we investigate different options that can be taken while using syntactical parsing and the effects of these choices on the lexical similarities and the properties of networks based on these similarities. The rationale here is that as we are using the text mining to map documents in order to identify clusters of fields or emerging topics we have to limit the textual information that we use to those elements in texts (or - more formally - those parts of speech) that actually contain the topics. Nouns or noun phrases are used as subjects, objects, predicative expressions or prepositions in sentences. Syntactic parsing as one of the applications within NLP will be used to extract the noun phrases from the abstracts; other categories, such as verbs, adjectives or adpositional phrases will be neglected. However, the selected noun phrase might contain an embedded phrase of these other types or even other embedded noun phrases. In what follows we first describe the data set that is used for the development of the methodology. Then, a short introduction to NLP is given, however, we refer to existing literature on this topic for an in-depth discussion.

Data source and data processing

A set of 6144 publications on 'Information Systems' is used in this study. We selected information systems as a field of study because it is a subject which "draws from various

reference disciplines, such as computer science, management, organisation studies, marketing, accounting, finance, economics, social psychology, sociology etc." (Cecez-Kecmanovic, 2002, 1699). While it "is not a young field of academic endeavo[u]r as it can look back at several decades of IS research" (Boell, 2012, 1), it can be viewed as an emerging discipline that still integrates knowledge from other reference disciplines (e.g. Currie and Galliers, 1999; Cecez-Kecmanovic, 2002). This makes it a fertile empirical setting for our analysis. This data set is retrieved from the *Social Sciences Citation Index* by using a custom developed search strategy focusing on 'Management Information System', 'Geographical Information System', 'Decision Support System' or 'Transaction Processing System' (Meyer et al., 2013). Publications from 1991 up to 2012 with document type Article, Letter, Note or Review are selected.

For the lexical component, the title and the abstract of the papers are processed by both Lucene (version 4.0) and the Stanford Parser. Terms used in the older single term based approach were retrieved by the next pre-processing steps: title and abstracts are merged and converted to lower case. Then, this data is tokenized by punctuation and white spaces. Stop words are removed through a custom built stop word list and remaining terms were stemmed by the Snowball Stemmer available in Lucene which is an extended version of the original Porter Stemmer (Porter, 1980). All terms that occur in only one document are removed. A term-by-document matrix is constructed in a vector space model with term frequency-inverse document frequency weightings (TF-IDF). Salton's cosine measure is used as similarity measure between documents (Salton & McGill, 1986).

Noun phrase extraction

For the extraction of noun-phrases we rely on the Stanford Parser, a Java package which has been developed and distributed by the Stanford Natural Language Processing Group. In short, this parser returns the grammatical structure of sentences based on probabilistic language models. In this study we use the PCFG-parser version 2.0.5 (Klein & Manning, 2003). The format of the output of the parser is a syntactic tree which describes the grammatical relations between words in a sentence (de Marneffe & Manning, 2008a, 2008b). In the output, nouns are tagged with NN or NNS (for plurals), noun phrases with NP. For detailed information on the parsing procedures and the resulting syntactic tree we refer to their website: http://nlp.stanford.edu.

Prior to the parsing, the DocumentPreprocessor, a Java Class provided in the Stanford Parser package, is used to extract separate sentences from the abstracts. Each sentence is numbered for retrieval purposes afterwards. Next, each sentence is parsed and each resulting noun phrase is numbered sequentially. Table 1 presents the output of the parsing of one sentence for one of the selected papers. We have added the labels.

```
Results of the study show that information systems downsizing may
produce benefits such as improved information systems, improved
organizational structure, higher productivity, and lower cost.
```

Table 1. Syntactic tree as a result of parsing the example sentence.

Label	Result of the Stanford Parser
	(ROOT
	(S
A	(NP
A1	(NP (NNS Results))
	(PP (IN of)

A2	(NP (DT the) (NN study))))
	(VP (VBP show)
	(SBAR (IN that)
	(S
B	(NP (NN information) (NNS systems) (NN downsizing))
	(VP (MD may)
	(VP (VB produce)
C	(NP
C1	(NP (NNS benefits))
	(PP (JJ such) (IN as)
C2	(NP
C2a	(NP (VBN improved) (NN information) (NNS systems))
	(, ,)
C2b	(NP (VBN improved) (JJ organizational) (NN structure))
	(, ,)
C2c	(NP (JJR higher) (NN productivity))
	(, ,)
	(CC and)
C2d	(NP (JJR lower) (NN cost)))))))))

There are several additional choices to be made for the processing and extraction of the noun phrases. It is the objective of this paper to study the consequences of these options. Each option or scenario is tagged differently.

For the retrieval of noun phrases from the parsed sentence, we have two options: Complete noun phrases (NP) can be selected or only leaf noun phrases in which no further noun phrases are embedded. Noun phrase *A* (`Result of the study`) in the example is such a complete noun phrase. But *A1* is a leaf noun phrase, it has no embedded NPs anymore. Analogously *C*, *C1, C2, C2a, C2b, C2c* and *C2d* are complete noun phrases while only *C1, C2a, C2b C2c* and *C2d* have no embedded NP. In this paper we will use the tag '`NP`' to denote complete noun phrases and `NPwoNP` for noun phrases without embedded noun phrases.

Next, the Snowball stemmer is applied on all terms in a noun phrase and stop words are removed. The stemmed terms within a single phrase are then used to create term-shingles. A term shingle is a set of subsequent terms. The length of these shingles can vary between 1 and the number of terms in the phrase which is the maximum. Table 2 presents the shingles that can be identified for noun phrases *A1* and *B* in the example. The number of terms in the phrases are 1 and 3 and that is also the length of the longest possible shingle in this phrase.

Table 2. All possible shingles extracted from noun phrases A1 and B

Noun Phrase	Code	Shingle	Length
A1: Results			
	A1a	`Results`	1
B: Information Systems Downsizing			
	Ba	`Information`	1
	Bb	`Systems`	1
	Bc	`Downsizing`	1
	Bd	`Information Systems`	2
	Be	`Systems Downsizing`	2
	Bf	`Information Systems Downsizing`	3

For the selection of shingles we identified five different possibilities with different criteria based on the number of terms in the noun phrase and on the length of the shingle. Each

possibility is labelled by an appropriate tag. Table 3 lists these five tags together with the criteria and their application on noun phrases A1 and B from the example.

Table 3. Five different selections of possible shingles extracted from noun phrases A1 and B

Tag	Criteria	A1	B
(none)	All possible shingles are included	A1a	Ba, Bb, Bc, Bd, Be, Bf
`Lm`	Only shingles with a length equal to the longest possible shingle are selected. length = maximum.	A1a	Bf
`lm_l1`	Only shingles with a length equal to 1 or shingles with a length equal to the maximum are selected	A1a	Ba,Bb,Bc, Bf
`l>1`	Only shingles with length higher than 1 are selected	(none)	Bd,Be,Bf
`m1_l>1`	Only shingles with length higher than 1 or shingles from single term noun phrases are selected.	A1a	Bd,Be,Bf

Next, Shingles can be recorded with the constituent words in the given order or the included terms can be sorted alphabetically. In our data set we found both following phrases 'information management system' and 'management information system'. Papers using these different versions of the same concept would not be linked to each other. As mentioned in the introduction a bi-gram detection method like the Dunning likelikehood ratio test would not be able to detect these variation. For each selected shingle, a alphabetically ordered version is created, the tags for the sorted scenarios get the suffix '_o'.

The combination of these five possible selection criteria with the options for the type of noun phrase and the possible sorting creates twenty different scenarios for the creation of a phrase by document matrix. This matrix contains only phrases or shingles that occur in more than one document and the weighting is a slightly modified TF-IDF version where the term frequency is equal to the number of sentences in which the phrase or shingle appears and IDF refers to the inverse of the number of distinct documents in which a shingle appears. Salton's cosine is calculated to express the similarity between documents. As a result we have for each document pair up to twenty different similarities based on the different scenarios in this NLP approach.

Bibliographic Coupling

For the citation component we calculate a bibliographic coupling cosine similarity based on the unique reference codes that Thomson-Reuters provides in its custom dataset. These codes are assigned to references in indexed papers and allow identification of common references between indexed documents without the requirement that also the cited document is indexed. This choice improves the application of bibliographic coupling in those fields where many cited documents are not indexed.

Hybrid Approach

The two components lexical and bibliographic coupling are combined by calculating a hybrid similarity as the cosine of the weighted linear combination of the underlying angles of each of the cosine similarities.

$$r = \cos\left(\lambda \cdot \arccos(\eta) + (1-\lambda) \cdot \arccos(\xi)\right), \quad \lambda \in [0,1], \tag{1}$$

where η is the similarity defined on bibliographic coupling and ξ the textual similarity. The λ parameter defines the *convex combination*, *arccos(η) and arccos(ξ)*, respectively, denote the two underlying angles (Glänzel & Thijs, 2012). For document pairs where one of the similarities is missing *arccos(0)* is uses as the underlying angle of this component. Document pairs where both similarities are missing are discarded. The effect of 7 different values of the λ parameter (0.125, 0.25, 0.33, 0.5, 0.66, 0.75 and 0.875) for combining the link-based similarity with the bestNLP based lexical component will be tested.

Clustering

Clustering of the data is done by the Pajek 'Single Refinement' implementation (Batagelj & Mrvar, 2003) of the Louvain method for community detection (Blondel et al., 2008). Prior to this clustering all singletons are removed from the network. The resolution parameter is set to 1 and five random restarts are requested. We report the number of clusters and the modularity of the clustering of each of the twenty nine networks that are created in data preparation (1 Single Term, 1 bibliographic coupling, 20 NLP versions and 7 hybrid networks).

Results

As a kind of benchmark, the results of the single term and the bibliographic coupling network are presented in Table 4. This table illustrates the problem of different distributional characteristics already mentioned by Janssens (2005). The bibliographic coupling network is very sparse with a density of only 6.6%. The Single term network is nearly complete with a density of 97.5% and average degree close to the number of nodes in the network (5991.8 vs. 6144). In the bibliographic coupling network 392 documents are singletons without any link with other documents, in the single term network there are only 3. These singletons are removed for the calculation of the weighted degree and the clustering.

Table 4. Network properties and clustering results for single term and bibliographic coupling networks
[Data sourced from Thomson Reuters Web of Knowledge]

		Average		Weighted Degree			Community Detection		
	Density	Degree	Single	Average	Med.	Max.	NC	Mod.	<10
BibC	6.6%	403.6	392	14.6	10.7	94.6	16	0.35	8
SingleT	97.5%	5991.8	3	217.2	219.0	392.5	5	0.04	1

The comparison of the weighted degree distribution (see Figure 1) proves once more the different nature of both networks.

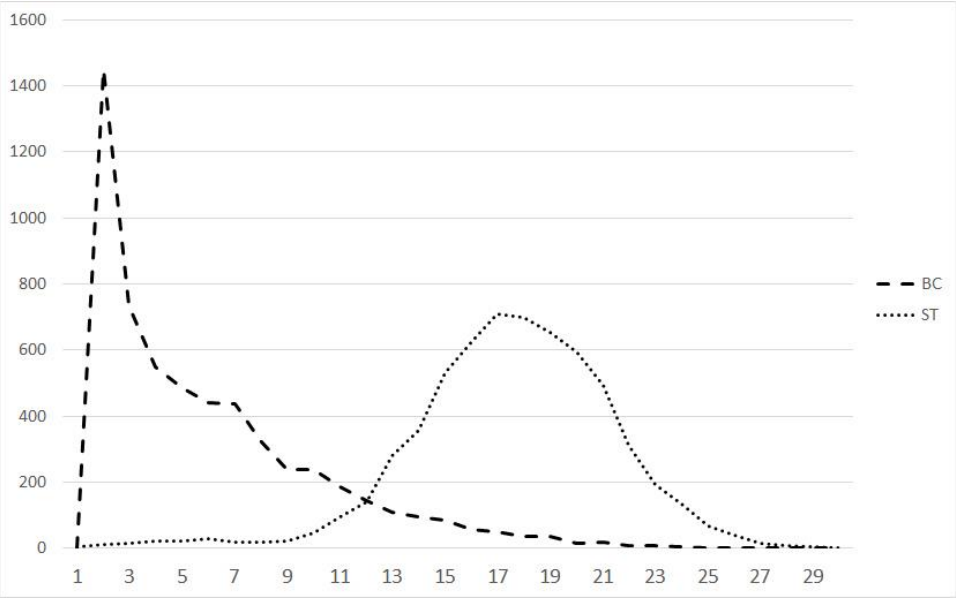

Figure 1. Distribution of weighted degree for the Bibliographic Coupling (BC) and the Single Term (ST) network.
[Data sourced from Thomson Reuters Web of Knowledge]

Consequently, also substantial differences in the results from the Louvain clustering are observed. In the `BibC` network, 16 clusters are found with a modularity of 0.35. However, eight small clusters with less than 10 documents are present. The modularity of the clustering of the `SingleT` network is extremely low, only 0.04 which means that despite the four large clusters, it is not possible to detect clear distinct topics in this network.
Another problem arises from the near completeness of the Singe Term network. The number of possible undirected links between documents grows quadratic with the number of nodes in a network and is given by the formula $\frac{n(n-1)}{2}$ where *n* denotes the number of documents. With this dataset of 6144 documents we already have 18.4 million lexical links. Using this approach comes with computational challenges for large datasets.

Comparing NLP scenarios

Next we compare the twenty NLP-based scenarios and their networks in Table 5. Due to the fact that in the complete NP approach the length of the noun phrases is higher than in the approach without embedded NPs, the files after parsing for NP contained many more lines and were about 50% larger in size. This is especially the case when comparing the two options without any restriction on length of the retrieved shingles. Ordering the phrases has no effect on number of returned lines or file size. For each scenario we calculated the total number of unique phrases found in the document set, the average number of unique phrases per document and the highest number of phrases in any document. After the construction of networks, also density and average degree are calculated. As we have the number of unique phrases and average number of phrases per document we could compute a density in a complete random network. In such a random network a number of phrases equal to the observed average number of phrases is assigned to each document. Possible links between documents can then be calculated and subsequently also the density in this completely random network. We used this formula to approximate the random density:

$$1 - \left(1 - \frac{C(u,a) - C(u-1,a)}{C(u,a)}\right)^{a} \quad (2)$$

where *u* denotes the total number of unique phrases, *a* the average number of phrases and *C(u,a)* denotes the number of combinations of set *u* with *a* elements. This random density provides us with a reference value for gauging the added information in the observed network. Table 5 reports this random density together with the ratio between the observed and the random density.

Several observations can already be made from the results presented in Table 5. The selection of type of noun phrase has an influence on the set of unique phrases and on the average and maximum number of phrases in the documents. As also observed with the file sizes, processing the complete noun phrases generates more data but when looking at the degree and density we don't see a large differences in network properties. When we apply sorting of the terms inside the phrases we detect a slight increase in the number of phrases that are included. Some of these sorted phrases are not excluded anymore as they appear on more than one paper. We find only very small differences in average degree and density between sorted and unsorted scenarios.

Table 5. Results of noun phrase extraction and network properties for twenty scenarios and the single term approach.
[Data sourced from Thomson Reuters Web of Knowledge]

	Phrases			Average		Random	Observed/
	Unique	Average	Maximum	Degree	Density	Density	Random
NP	33433	73.7	239	5908.41	96.2%	15.1%	6.4
NPo	34335	74.7	242	5908.41	96.2%	15.1%	6.4
NPwoNP	26978	63.6	201	5870.08	95.6%	14.1%	6.8
NPwoNP_o	27323	64.2	203	5870.08	95.6%	13.9%	6.8
NP_lm_l1	19078	62.0	200	5908.41	96.2%	18.3%	5.3
NP_lm_l1_o	19351	62.2	202	5908.41	96.2%	18.0%	5.3
NPwoNP_lm_l1	17762	55.1	174	5870.08	95.6%	15.7%	6.1
NPwoNP_lm_l1_o	17932	55.3	175	5870.08	95.6%	15.5%	6.1
NP_lm	16080	25.5	94	3546.63	57.7%	3.8%	15.1
NP_lm_o	16353	25.8	96	3548.91	57.8%	4.1%	14.2
NPwoNP_lm	15043	24.6	91	3520.17	57.3%	4.1%	14.1
NPwoNP_lm_o	15213	24.8	93	3522.20	57.3%	4.0%	14.2
NP_m1_l>1	30435	37.2	111	3983.86	64.9%	4.4%	14.7
NP_m1_l>1_o	31338	38.2	115	3996.44	65.1%	4.5%	14.4
NPwoNP_m1_l>1	24259	33.2	103	3919.09	63.8%	4.4%	14.5
NPwoNP_m1_l>1_o	24604	33.7	106	3927.19	63.9%	4.6%	13.9
NP_l>1	27237	21.6	85	1475.91	24.0%	1.8%	13.6
NP_l>1_o	28151	22.6	82	1506.60	24.5%	1.9%	13.2
NPwoNP_l>1	21147	17.9	64	1350.85	22.0%	1.5%	14.5
NPwoNP_l>1_o	21493	18.4	64	1370.96	22.3%	1.5%	14.9
SingleT	6891	68.2	236	5991.76	97.5%	49.1%	2.0

The largest effect comes from the selection of length of shingles. When no additional criteria are applies we observe an average degree and density close to the properties of the single term

approach. Also when we select in each phrase only the individual terms complemented with the complete phrase (tagged with _lm_l1) we obtain similar results. The construction of these type of networks will suffer from the same computational complexity as the single term network. Next, the four scenarios where only the complete phrase is retained (_lm) have the lowest number of unique phrases but still have a density slightly above 57%. The fourth set of scenarios where the single term phrases and all shingles with length more than one are selected, have large sets of unique phrases and a density which is a bit higher than the third set of scenarios. Both the latter sets have density ratios between 13.9 and 15.1. It is clear that removing shingles with size 1 from longer phrases has a large effect on the density of the network. This effect is even more pronounced in the last set of scenarios where all shingles with size one are removed even when the noun phrase consists only of one term. This last criteria implies that many phrases are neglected. This is also reflected by the average number of phrases in the documents. Now the density also drops below 25%.

Next singletons are removed from each of these networks, weighted degree is calculated and the Louvain community detection is used to cluster the documents. The results can be found in table 6. First we observe that the selection of type of noun phrases or the sorting of terms has no effect on any of the results in this table. From the difference between average and median of the weighted degree and the ratio between average and the maximum we learn that the first two sets of scenarios (no restriction on length of shingles or all one term singles complemented with the complete phrase) have a distribution that is close to normal just as the single term approach. Also the modularity coefficient is low with values around 0.1

Table 6. Weighted Degree of reduced network and Louvain Community Detection result
[Data sourced from Thomson Reuters Web of Knowledge]

		Weighted Degree			Community Detection		
	Singletons	Average	Median	Max	NC	Mod.	<10
NP	5	86.59	86.85	168.07	7	0.102	1
NPo	5	85.35	85.60	167.04	8	0.102	0
NPwoNP	5	88.08	88.31	174.58	8	0.107	2
NPwoNP_o	5	87.38	87.63	166.82	8	0.108	2
NP_lm_l1	5	113.73	113.95	233.21	7	0.092	1
NP_lm_l1_o	5	112.83	113.04	233.26	6	0.920	0
NPwoNP_lm_l1	5	110.35	110.76	227.28	8	0.098	2
NPwoNP_lm_l1_o	5	109.64	110.13	227.08	7	0.097	2
NP_lm	24	47.44	46.29	123.49	8	0.107	1
NP_lm_o	24	46.81	45.56	122.78	7	0.107	1
NPwoNP_lm	24	48.69	47.27	140.19	8	0.108	1
NPwoNP_lm_o	24	48.18	46.76	139.71	8	0.108	1
NP_m1_l>1	13	34.23	33.32	101.48	10	0.147	2
NP_m1_l>1_o	13	33.79	32.86	100.13	9	0.147	2
NPwoNP_m1_l>1	13	37.46	36.42	110.33	10	0.147	2
NPwoNP_m1_l>1_o	13	37.24	36.19	109.51	10	0.143	2
NP_l>1	35	15.12	13.60	105.81	12	0.341	2
NP_l>1_o	34	15.32	13.81	103.87	12	0.330	2
NPwoNP_l>1	38	16.39	14.62	119.63	14	0.348	2
NPwoNP_l>1_o	38	16.64	14.86	118.50	12	0.338	2

SingleT	3	217.16	219.03	392.51	5	0.043	1

In the next two sets of scenarios one can observe that the distribution of the weighted degree deviates from a normal distribution. The scenarios where only the complete phrase is has also a low modularity. When we start to include all shingles with a length higher than one the clustering performs better. In the last set of scenarios where also the single term phrases are excluded we find the highest modularity and weighted degree distribution that is close to the distribution found in bibliographic coupling.
Based on these findings and together with the reduced data size for restricted noun phrase we chose the scenario with this tag: NPwoNP_l>1_o as the best possible approach. The sorted version is selected as it includes slightly more phrases than the unsorted version and solves the problem of multiple variation of the same concept.

Hybrid Combination

In this section we compare seven hybrid combinations of the bibliographic coupling component together with the selected NLP component. The different weights for the components are (0.125, 0.25, 0.33, 0.5, 0.66, 0.75 and 0.875). In Table 7, the top and bottom row contain the results of the two components for reference. After the hybrid combination, 25 documents remained singletons in the network and were removed.

Table 7: Results of hybrid clustering with different weight parameters
[Data sourced from Thomson Reuters Web of Knowledge]

	Weighted Degree			Community Detection		
Weight λ	Average	Median	Max	NC	Mod.	<10
NPwoNP_l>1_o (λ= 0)	16.64	14.86	118.50	12	0.338	2
0.125	16.26	14.88	104.66	12	0.322	2
0.25	15.90	14.82	90.66	11	0.312	2
0.33	15.68	14.58	81.62	11	0.308	2
0.5	15.19	13.64	62.22	10	0.310	3
0.66	14.73	12.40	69.24	10	0.317	3
0.75	14.47	11.62	75.95	10	0.323	4
0.875	14.11	10.48	85.27	10	0.333	3
BibC (λ= 1)	14.62	10.71	94.59	16	0.350	8

We would like to recall that the appropriate choice of the weight parameter λ used to be crucial for the quality of the clustering result with a possible distortion of the results by too much weight on the single term lexical approach (Janssens et al. 2008). However, Table 7 clearly shows that the distribution of weighted degree is not distorted by any choice of the λ parameter. Also, for each of the chosen values a modularity above 0.3 is obtained. And finally, when looking at the number of clusters, it evolves from 12 in the top row (lexical component) to 10 in the λ=0.5 weighting scheme to 16 in the bottom rown (link component).
When we look at the correspondence of cluster assignment between two schemes we observe higher stability between schemes with λ values closer to each other. Cramer's V measures are calculated between all schemes and plotted in Table 8.

Table 8.Cramer's V measurement of association
[Data sourced from Thomson Reuters Web of Knowledge]

	Lexical	0.125	0.25	0.33	0.5	0.66	0.75	0.875
0.125	0.85							
0.25	0.79	0.86						
0.33	0.76	0.80	0.89					
0.5	0.66	0.71	0.74	0.74				
0.66	0.63	0.65	0.68	0.71	0.90			
0.75	0.62	0.64	0.66	0.69	0.87	0.93		
0.875	0.59	0.61	0.63	0.66	0.84	0.93	0.91	
BibCpl	0.30	0.33	0.40	0.44	0.65	0.70	0.77	0.75

Application

This section outlines briefly the results of our partitioning of the hybrid network (with 50% weight on both components) at three levels with increasing resolution (level I with resolution of 0.7; II with 1.0 and III with resolution 1.5). As mentioned above, we used a data set of 6144 publications in Information System Research for our analyses. Level I resulted in three large clusters and two pairs or triplets of papers with no link to any other documents. These pairs/triplets (five papers at level I) are removed from further analysis. At level II we found seven clusters and three pairs/triplets (8 papers) and level III has 19 clusters and the same 8 papers were grouped in three pairs/triplets. These findings are also summarized in figure 2. Although the three levels consist of independent runs of the Louvain cluster algorithm we can observe a near-perfect hierarchical structure. This is confirmed by Cramér's-V values of 0.94 between level I and II, 0.93 between I and III and 0.84 between levels II and III. The labels of each cluster at the three levels are taken from the titles of core documents within each cluster. These core documents have been determined according to Glänzel & Thijs (2011) and Glänzel (2012) on the basis of the *degree h-index* of the hybrid document network. In particular, core documents are represented by core nodes which, in turn, are defined as nodes with at least *h* degrees each, where *h* is the h-index of the underlying graph and only edges with a minimum weight of 0.15 are retained.At the lowest level the three clusters contain publications that fit in broad categories, such as '*planning/development/ implementation'* (cluster I.2 with 3855 papers), '*user and technology acceptance'* (cluster I.3 with 1302 papers), and '*decision support systems'* (cluster I.1 with 957 papers).

Given the size of the *planning/development/implementation* cluster and the hierarchical structure of the different levels, there is value in exploring the clustering at a higher resolution which allows us to develop a more differentiated understanding of the IS literature that falls in this category. At Level II with a resolution of 1.0 we identify 5 clusters. There are three large clusters: '*II.c strategic IS planning*' (1414 papers), '*II .b development /OSS /planning'* (1119 papers), '*II.e supply cha*in' (1108). Smaller clusters were also found with one midsized cluster: '*II.f intangible assets*' (376) and one small but emergent topic: '*II.h security*' (48). This last cluster is not further partitioned at level III. The three large Level II clusters can be divided further. At a resolution of 1.5 the following picture emerges:

- *Strategic IS planning* with 1414 papers: can be subdivided in two large clusters on *strategic planning* (III.5) and *performance measurement* (III.8); 2 midsized clusters on *HR & Accounting* (III.11) and *ERP* (III.18); and a small cluster containing *executive perspectives* (III.13).
- *Development/OSS/Planning* with 1119 papers: The cluster *'Design science in IS research*' accounts for more than half of the papers and focuses on IT implementation

and methods (III.12); another large cluster is centred on *Systems development projects* (III.2); a smaller cluster on *conceptual models* (III.15), with close to 50 papers, contains also bibliometric studies on, e.g., 'citation classics'.

- *Supply chain*: As one would expect the largest grouping is associated with *Supply Chain Management* (III.10), followed by *firm performance* (III.14) and *open source* (III.19); a small but still substantial cluster focuses on *outsourcing* (III.9), another small cluster can be linked to *innovation, assimilation and diffusion* (III.6).

The midsized Level II cluster on *intangible assets* can be further split up into *knowledge management* (III.16) accounting for most of the papers (367) and *customer relationship management* (CRM, III.22, 24 papers).

Clustering for the other Level I areas (*decision support systems*, *user and technology acceptance*) remains broadly the same. Only at level III with a resolution set at 1.5 can one identify a finer grained structure for these clusters:

- *Decision support systems* (cluster III.1) can be further differentiated from a cluster (III.7) around *communication and virtual teams*;
- *User satisfaction and service quality* (cluster III.3) are differentiated from *usage and acceptance of technology* (cluster III.4).

Level I Resolution = 0.7			Level II Resolution = 1.0			Level III Resolution = 1.5		
Cluster	# Pubs	Label	Cluster	# Pubs	Label	Cluster	# Pubs	Label
1	957	Decision Support Systems	a	955	Decision Support Systems	1	807	Decision Support System
						7	111	Communication, Virtual Teams
2	3855	Development, Implementation, Planning	b	1119	Development, (Open source) Software, Planning	2	398	Development Projects
						12	537	Design Science in IS Research
						15	43	Conceptual Modeling
			c	1414	Strategic IS Planning, Management	5	454	Strategic Planning
						8	429	Performance Measurement
						13	51	Executive Perspective
						11	359	HRM & Accounting
						18	210	Enterprise Resource Planning
			e	1108	Supply Chain	6	24	Innovation, Assimilation & Diffusion
						9	107	Outsourcing
						14	338	Firm Performance
						10	442	Supply Chain Management
						19	334	Open Source
			f	376	Intangible Asset Management	16	367	Knowledge Management
						22	26	Customer Relationship Management
			h	48	Security	20	78	Security
3	1302	User Oriented	d	1092	Satisfaction, Technology Acceptance Model	3	288	Satisfaction, Service Quality
						4	709	Technology Use and Acceptance

Figure 2.Cluster solution at three levels with increasing resolution

[Data sourced from Thomson Reuters Web of Knowledge]

Conclusions

Based on the data presented in this paper we can conclude that the extraction of noun phrases from abstracts and titles can improve the lexical component in the hybrid clustering. However, using the noun phrase itself is not sufficient for the improvement. The scenario where the complete retrieved noun phrase is used only reduces the density of the network but does not have an effect on the clustering afterwards. Only when the data is restricted to shingles with at least two terms constructed out of the noun phrases an improvement in the clustering is observed. This solution has several advantages over the other scenarios. It has a small set of unique terms, the density of the network is reduced to a quarter of the network constructed on single terms and the distribution of weighted degree is similar to the distribution in bibliographic coupling. As a consequence the risk of distorting the network by choosing not the optimum parameter or even an inappropriate parameter in the hybrid approach is distinctly reduced. It seems that the parameter will not be used anymore in a function to set the right focus on the document set but to change the viewpoint while the clustering stays in focus.
We even found out that many of the shingles only appear once in each document which allows us to bring the calculation of similarities in the lexical approach more in line with the bibliographic coupling by abandoning the TF-IDF weighting and adopting a binary approach.

Acknowledgement

This paper is an extended version of a research in progress paper presented at 'Mining Scientific Papers: Computational Linguistics and Bibliometrics' workshop held at the 15th International Conference on Scientometrics and Informetrics, Istanbul, Turkey, June 29th – July 3rd.

References

Batagelj, V. & Mrvar, A. (2003). Pajek–Analysis and visualization of large networks. In M. Jünger & P. Mutzel (Eds.), *Graph drawing software* (pp. 77–103). Berlin: Springer.

Blondel, V.D., Guillaume, JL., Lambiotte, R. & Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, P10008

Braam, R.R., Moed, H.F., van Raan, A.F.J. (1991a). Mapping of science by combined cocitation and word analysis, Part 1: Structural aspects. *JASIS*, 42 (4), 233–251.

Braam, R.R., Moed, H.F., van Raan, A.F.J. (1991b). Mapping of science by combined cocitation and word analysis, Part II: Dynamical aspects. *JASIS*, 42 (4), 252–266.

Boell, S.K. (2012). *Theorizing Information and Information Systems*. Doctoral thesis. Sydney: University of New South Wales.

Callon, M., Courtial, J. P., Turner, W., & Brain, S. (1983). From translations to problematic networks. An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235

Cecez-Kecmanovic, D. (2002). The discipline of information systems: issues and challenges. In *Proc. of the 8th Americas Conference on Information Systems*, pp. 1696-1702.

Currie, W. & Galliers, B. (Eds.) (1999). *Rethinking Management Information Systems: an interdisciplinary perspective*. New York: Oxford University Press.

de Marneffe, M.C. & Manning, C.D. (2008). *Stanford Dependencies manual*.

de Marneffe, M.C. & Manning, C.D. (2008). The Stanford typed dependencies representation. In *COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Dunning, T. (1993) Accurate methods for the statics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Glänzel, W. & Thijs, B. (2011), Using `core documents' for the representation of clusters and topics. *Scientometrics*, 88 (1), 297–309.

Glänzel, W. & Thijs, B. (2012). Using `core documents` for detecting and labelling new emerging topics. *Scientometrics*, 91(2, 399-416.
Glänzel, W. (2012), The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, 93 (1), 113–123.
Glenisson, P., Glänzel, W., Janssens, F & De Moor, B. (2005) Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management*, 41, 1548-1572.
Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1), 163–180.
Jain A., Dubes R. (1988), *Algorithms for clustering data*, Prentice Hall.
Janssens, F., Glenisson, P., Glänzel, W. & De Moor, B. (2005) Co-clustering approaches to integrate lexical and bibliographical information. In *Proc. of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* p. 284-289, Stockholm: Karolinska Univerity Press
Janssens, F., Leta, J., Glänzel, W. & De Moor, B. (2006) Towards mapping library and information science. *Information Processing and Management*, 42, 1614-1642.
Janssens, F., Glänzel, W. & De Moor, B (2008) A hybrid mapping of information science. *Scientometrics*, 75 (3), 607-631.
Klein, D. & Manning, C.D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.
Lucene.apache.org. (2015). *Apache Lucene*. [online] Available at: http://lucene.apache.org [Accessed 4 January 2015].
Maning, C.D. & Schütze, H. (2000), *Foundations of Statistical Natural Language Processing*. Combridge: MIT press.
Meyer, M., Grant, K., Thijs, B., Zhang, L., Glänzel, W. (2013), *The Evolution of Information Systems as a Research Field*. Paper presented at the 9th International Conference on Webometrics, Informetrics and Scientometrics and 14th COLLNET Meeting, August 15-17, 2013 Tartu, Estonia.
Noyons, E. C. M., & Van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an R&D field. The case of Optomechatronics. *Scientometrics*, 30, 157–173.
Porter, M.F., (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.
Salton, G. & McGill, M.J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

FACULTY OF ECONOMICS AND BUSINESS
DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 67 00
fax + 32 16 32 67 32
info@econ.kuleuven.be
www.econ.kuleuven.be/MSI